Rejuvenating multiple true–false: Proposing fairer scoring methods

Thomas Puthiaparampil, MD, Md Mizanur Rahman, PhD, Sabrina Binti Lukas, MMed, Nariman Singmame, MEmMed, Shazrina Binti Ahmad Razali, MSc Medical Education

Faculty of Medicine and Health Sciences, Universiti Malaysia Sarawak, Sarawak, Malaysia

ABSTRACT

Introduction: Multiple true–false tests (MTF) with penalty scoring consistently delivered low scores and many failures for over two decades in our medical faculty. This issue remained unaddressed, as the overall student performance was redeemed by other assessments like Best Answer Questions and Modified Essay Questions. The post-test item analyses revealed that there were several items with unacceptable difficulty index and discrimination index, many omissions, and that the false options performed worse than the true options in the difficulty index but better in the discrimination index. This study aimed to evaluate some final professional examination MTF papers to propose possible remedial measures.

Materials and Methods: We examined 5 years' final professional examination MTF results, their item analysis, the student performance in true and false items and failure rates.We explored the impact of excluding the flawed questions post-test based on item analysis and redoing the scores. We also explored the effect of removing the penalty scoring and recalculating the scores.

Results: The two new scoring methods, such as post-weeding recalculation and no-penalty proportionate scoring, showed remarkable improvement in scores and also reduced the failure rates significantly compared to the penalty-scoring model.

Conclusion: We propose two new scoring methods for MTF, which would be fairer to the students and would have the prospect of rejuvenating MTF tests.

KEYWORDS:
MTF with no penalty scoring, multiple true–false, post-test weeding, MTF scoring methods

INTRODUCTION

The American National Board examinations scrapped Multiple True–False (MTF) tests, as it was not considered suitable to test higher knowledge than recall of facts.[1] However, this notion was disputed.[2] MTF is still used widely in other parts of the world.[2-5] MTF was considered superior to One Best Answer Questions (OBA or BAQ), because it could test five individual items of knowledge in one question.[2-4,6-8] Other advantages attributed to MTF were it could accommodate complex scenarios[6]; it could test minute understanding of students[9] and it allowed extensive feedback to students that could stimulate learning.[3,10] Factual knowledge is essential for a doctor to function efficiently, and MTF is the best at testing it.[2,3,7,11-13] The ability to discriminate false statements from true, which is required of doctors, is well-tested in MTF.[2] Omission is not an option in real-life clinical situations, and so, it should not be allowed in MTF also.[14] Omission is eliminated in the no-penalty or number-right scoring model. Most criticisms against MTF could be traced to attempts at testing higher knowledge than recall of facts, flaws in the questions and inadequate vetting of questions.[2,3,7,10] Construction of flawless MTF questions, thorough multidisciplinary vetting, post-test analysis and feedback to the question authors were considered very important for quality assurance.[1,6,15] Getting feedback from the examinees about the questions could be a valuable measure to improve the standard of MTF.[3] MTF performance should correlate well with the performance of other theory assessments, and poorly performing questions should be dropped from the question bank.[1,6,16] OBA also is not free from the guessing issue, and it might overestimate the students' knowledge.[9] Some studies reported the poor correlation of MTF with other theory assessments and the better performance of OBA.[10,15]

How to score MTF is a disputed issue. Kanzow et al.[6] described over a dozen penalty scoring algorithms, each of which produced different scores on the same test and concluded that none of them was worth recommending. Similarly, Schmidt et al.[14] described over two dozen scoring methods. However, none of them was shown as universally conclusive. Some studies advocated the abolition of the penalty scoring[12,13], while some advocated keeping it.[3,5,7] MTF without penalty scoring would reduce score variability and attenuate discrimination between examinees.[13] Many universities have adopted the no-penalty scoring system, where the correct responses are given points, and the incorrect responses and omissions are ignored.[12,13] The possibility of scoring at least 50% by blindly answering all the items as true in the no-penalty model remains unresolved.[13,17] The penalty scoring led to many items being left unattempted and low scores in MTF in our institution. The same pattern was repeated in almost all minor and major examinations of the faculty for over two decades. In one of our previous studies, we argued that the inherent flaws in MTF could not be remedied, as fewer false (F) options were answered correctly and omitted

This article was accepted: 26 February 2023
Corresponding Author: Thomas Puthiaparampil
Email: pthomas@unimas.my

more often than true (T) options.[17] Good F options were harder to construct, but they had a better discrimination index, meaning more higher performing students answered them correctly.[16,17] Since MTF with and without penalty scoring have unresolved issues, we explored ways to uplift both of them.

This study was triggered by the observation of the consistent poor performance of medical students in MTF papers. Our faculty used penalty scoring in the 5-option MTF tests in which each correct answer got 1 point, each incorrect answer got –1 point and '0' point for omission. The negative marks were not carried over from one question to the next. The poor performance in MTF was attributed to the penalty scoring. Furthermore, MTF was always used along with BAQ and MEQ, which covered up the issue. Our previous studies revealed that the flaws in the questions, especially the careless construction of false items, contributed to this issue,[17] and that the MTF performance in the final professional examination (FPE) adversely impacted the final scores and grades of the graduates.[18] In this context, we explored the feasibility of rejuvenating MTF with new scoring methods for penalty-MTF and no-penalty MTF, which would consider the flaws in the tests and also make MTF fairer for the students.

MATERIALS AND METHODS

This study was conducted in a public university in Malaysia with formal approval by the faculty's dean and the ethics committee of the university. We examinedthe data from five FPEs (A,B,C,D,E),which used 60 five-option MTF questions as one of the three theory papers.

Data Preparation

The original penalty-MTF scores were noted from five FPE results. The distribution of T and F items was noted. Students' optical mark reader (OMR) reports were checked to get the number of T items and F items answered correctly. The total of these served as the no-penalty scores. The number of omissions in each question was also noted.We studied the pass/fail rates in the three sets of scores. The three sets of scores obtained by the three scoring methods were compared and statistically analysed.

The three scoring methods and sets of scores we compared were:

1. The original one with penalty scoring, as practised in the faculty
2. The post-weed: scores recalculated after weeding the flawed questions from the original (recalculation was done by the OMR machine). Flawed questions include (a) those incorrectly answered by 60% or more students (difficulty index (DIFI) of <0.4); (b) questions with 40% or more omissions; (c) those with 0 or negative discrimination index (DISI)
3. With no-penalty scoring: the scores were noted from the OMR reports, which provided the T and F items answered correctly.

The pass score for all the sets was 50%. The no-penalty set had an additional criterion, which aimed to offset the possibility of scoring by blind guessing in future tests: there should be a minimum score of 20% from F items and 20% from T items. If either F or T score was less than 20%, for each two correct F, 3 T would be counted. If both the F and T scores were 20% or more, all correct F and T items would be counted.

Data Analysis

All the data were captured in Microsoft Excel and then transferred to IBM SPSS for analysis. The mean percentage scores of the original MTF tests, post-weed scores and no-penalty scores were compared with a dependent (paired) sample t-test. This test aimed to examine the mean difference between the original scores versus the post-weed scores and the original scores versus the no-penalty scores. We calculated the Cohend to examine the practical significance (effect size). Apart from this, we also categorised the scores into 'pass' and 'fail' of the three sets. A non-parametric Cochrane Q test was done to obtain the statistical difference among the three sets. A *p* value of less than 0.05 was considered statistically significant.

RESULTS

Table I illustrates the descriptive statistics of the students' original MTF scores with penalty, post-weed scores and scores with no penalty from the five FPEs. Data analysis revealed that the mean difference between the original and post-weed varied from 2.58 percentage points to 5.43 percentage points. The percentage score differed substantially between the original and the no-penalty category, which ranged from 11.36 percentage points to 14.04 percentage points. The year-wise paired sample t-test indicated a statistically significant difference in the penalty scores versus post-weed scores (*p*<0.001) with large Cohend. Similarly, a statistically significant difference was found between penalty scores versus no-penalty scores (*p*<0.001), and the effect size was large.

Table II illustrates the students' pass/fail rates resulting from the three scoring methods. The passing rate in MTF with penalty was very low. It varied from 15.2% to 28%. The passing rate improved with post-weed recalculation, which varied from 25.9% to 49.2%. In MTF without penalty, the passing rate was substantially higher. It varied from 70.8% to 89.3%.

Five hundred and eighty-five students' scores were examined to determine the pass rate changes with three scoring methods. Cochrane's Q test determined that there was a statistically significant difference in the proportion of students who passed the tests, $\chi^2(2) = 556.480$ (2), $p < 0.001$ (Figure 1). A post-doc pair-wise analysis revealed that there was a statistically significant difference between penalty MTF versus post-weed (test statistic=.142, *p*<0.001), similar to penalty MTF versus no-penalty MTF (test statistic=.592, *p*<0.001). The test also showed that there was a statistically significant difference between post-weed and no-penalty MTF (test statistic=.452, *p*<0.001).

Table III demonstrates the trends in T and F distribution and the students' performance in the five tests. The proportion of true items was more than false items in a 55:45 ratio; about 65% T items were answered correctly, while only about 46% F items were answered correctly. The omission rate varied from 24.1 to 29, with a mean of 26.8% (Table III).

Table I: Mean MTF scores obtained with three scoring methods

FPE	N	Scores with penalty				Mean difference	Cohen-d	No-penalty scores		Mean difference	Cohen-d
		Original	-	Post-weed							
		Mean	SD	Mean	SD			Mean	SD		
A	112	41.79	7.48	44.37	8.02	2.58	2.32	55.57	7.34	13.78	1.38
B	118	45.79	6.84	50.04	7.34	4.25	4.02	57.14	7.23	11.35	15.50
C	122	44.70	6.73	47.93	6.83	3.23	3.10	58.74	7.06	14.04	3.49
D	106	42.28	7.86	47.71	8.53	5.43	3.14	54.59	8.41	12.31	3.89
E	127	43.83	7.51	47.04	7.78	3.21	2.78	56.35	7.62	12.52	3.15

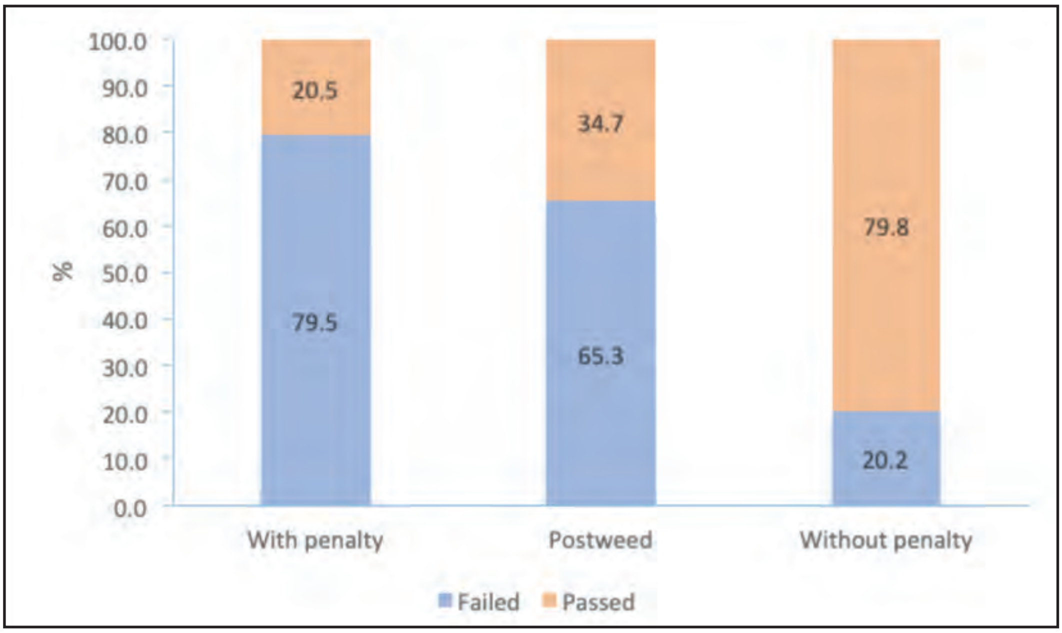
Statistical test obtained from paired sample t-test (Score with penalty vs Post-weed) and (Score with penalty vs score without penalty)
Cohen d = 0.2, 'small', 0.5 = 'medium' and >0.8 'large' effect size.

Table II: Pass/fail rates with 50% cut-off obtained with three scoring methods

FPE	With penalty Original				With penalty Post-weed				Without penalty			
	Fail		Pass		Fail		Pass		Fail		Pass	
	n	%	n	%	n	%	n	%	n	%	n	%
A	95	84.8	17	15.2	83	74.1	29	25.9	23	20.5	89	79.5
B	85	72.0	33	28.0	60	50.8	58	49.2	25	21.2	93	78.8
C	96	78.7	26	21.3	79	64.8	43	35.2	13	10.7	109	89.3
D	87	82.1	19	17.9	71	67.0	35	33.0	31	29.2	75	70.8
E	102	80.3	25	19.7	89	70.1	38	29.9	26	20.5	101	79.5

Table III: True and false item distribution and scoring in five MTF tests

FPE	Distribution		Omission %	Answered correctly	
	T (%)	F (%)	Mean and (Range)	T% (Range)	F% (Range)
A	54.7	45.3	27.2 (3–46.67)	64.6 (81–44)	44.7 (78–22)
B	54.0	46.0	26.5 (3.33–48)	63.7 (83–39)	49.5 (74–30)
C	53.3	46.7	24.1 (0–43.67)	67.9 (87–48)	48.0 (73–30)
D	56.3	43.7	29.0 (7.33–52.67)	62.8 (84–46)	44.1 (69–21)
E	58.7	41.3	27.4 (6.98–43.49)	64.9 (88–45)	44.2 (89–34)
Mean	55.4	44.6	26.8 (0–52.67)	64.8	46.1



Cochran's Q test (df)= 556.480 (2), p < .001

Fig. 1: Overall pass/fail rates in MTF with three scoring methods

DISCUSSION

Test reliability improves with test length, and MTF being easier to construct would make it possible to include more items, which would broaden the subject coverage.[19] Penalty scoring leads to omissions, which narrows the score distribution and lowers the test reliability.[19] With no-penalty scoring, the issue of guessing would be mitigated, and the validity and reliability of the test would improve by increasing the number of items.[20] MTF being easier to construct and allowing to test more facts than BAQ and extended matching question (EMQ), we find it worth rejuvenating it with new scoring methods, which would make it fairer to the students and viable to use no-penalty scoring. The issue of blind guessing seems to be ignored generally with no preventive measures suggested even in the 27 MTF scoring methods described in a systematic review article published in 2021.[20] In our setting, the general tendency has been to blame the students for their low MTF scores and ignore the quality of questions as a possible contributing factor. One of our previous studies discussed this issue.[16] The expert vetting would have passed the questions as 'perfect', but the post-test item analysis revealed the flaws in the questions. The rate of omission, DISI and DIFI were considered while recruiting questions for question bank,[16] as these indicators are considered valuable to judge the quality of the items. Standard error of measurement is lowered, and test reliability is reduced if the test contains very easy or very difficult items.[20] If some of the items were not suitable for further use, how could they be suitable for the current use? This concept led us to weed out flawed questions post-test and adjust the scores to benefit the students. We chose a DIFI of <0.4 and a DISI of ≤0 as cut-off points for exclusion of questions for score recalculation. In no-penalty model, guessing is permitted, and scores are higher as omission is eliminated.[20] Our results showed a consistent pattern of scores improving with the weeding of flawed questions and with no-penalty scoring (Tables I and II; Figure 1). Both of them showed the potential to rejuvenate the MTF tests.

MTF is the only test with penalty scoring. The fear of penalty leading to many omissions and the penalty-scoring leading to loss of scores were the apparent reasons for the poor scoring and the high failure rates in MTF. There is no reason for applying a penalty other than to prevent blind guessing. Moving to no-penalty scoring, we needed to devise an alternative method to preclude blind guessing, as students would know by experience that more than 50% of the items might be true. So, why not just answer all the items as true! In the absence of penalty, such a trick would secure as many scores as the number of T items in the paper. Discarding penalty scoring without any safeguards against blind guessing would be unwise. To surmount blind guessing, we have proposed a minimum score of 20% each for both F and T items and a proportionate scoring of T:F::3:2. This was based on our finding that in the five MTF papers the T and F ratio was approximately 55:45, and the mean of F items answered correctly was 46.1% (Table III). In the absence of 20% F scores or T scores, the score would be calculated in a proportion of 2:3::F:T. This is based on the proportion of correctly answered F and T items in five FPEs (Table III). If the F and T scores both exceed 20%, no restrictions would apply. In the 5 years' results, none of the students scored less than 20% in either F or T items (Table III). This could be explained as these scores were calculated without penalty, after the tests were done in the penalty scoring mode. Only when the faculty moves from penalty to no-penalty MTF, this proposal could be validated. It would also be wise to include Extended Matching Questions, Short Answer Questions or Very Short Answer Questions to broaden the assessment.

LIMITATIONS

We could not use authentic no-penalty MTF, as our faculty did not practise it yet. Therefore, the new scoring method for no-penalty MTF could not be validated. The no-penalty scores we used for this study were derived from original MTF tests with penalty. We removed the negative scores deducted as penalty to get the no-penalty scores. In this method, the scores could be slightly lower than the actual no-penalty MTF, as omission would be eliminated in no-penalty tests.

CONCLUSION

We are facing the prospect of a valuable assessment tool like MTF withering away, as the student scores are consistently low in these tests. It is worth rejuvenating MTF, as it has several pluses. We propose post-weed score recalculation for the penalty-scoring MTF and a minimum F and T passing score with a proportionate F and T scoring method for the no-penalty scoring MTF.

CONFLICTS OF INTEREST

None of the authors declared any conflict of interest.

ACKNOWLEDGEMENTS

The authors are thankful to the Universiti Malaysia Sarawak (UNIMAS) and the Dean of the Faculty of Medicine and Health Sciences for approving this project.

ETHICAL APPROVAL

Ethical approval was obtained from the Ethics Committee (UNIMAS/TNC(PI)/09-65/01(47) RUJUKAN ETIKA: FME/22/39 of Universiti Malaysia Sarawak (UNIMAS). We also obtained administrative approval from the dean of the Faculty of Medicine and Health Sciences.

FUNDING

No funding was obtained for this project.

REFERENCES

1. Richardson R. The multiple choice true/false question: what does it measure and what could it measure? Med Teach 1992; 14(2-3): 201-4.
2. Anderson J. For multiple choice questions. Med Teach 1979; 1(1): 37-42.

3. Biran LA. Hints for students (and examiners) on answering MCQ questions of the multiple true/false type. Med Teach 1986; 8(1): 41-8.
4. Mitchell G, Ford D, Prinz W. Optimising marks obtained in multiple choice question examinations. Med Teach 1986; 8(1): 49-53.
5. Lahner FM, Lörwald AC, Bauer D, Nouns ZM, Krebs R, Guttormsen S. Multiple true–false items: a comparison of scoring algorithms. Adv Health Sci Educ 2018; 23(3): 455-63.
6. Kanzow P, Schuelper N, Witt D, Wassmann T, Sennhenn-Kirchner S, Wiegand A. Effect of different scoring approaches upon credit assignment when using Multiple True-False items in dental undergraduate examinations. Eur J Dental Educ 2018; 22(4): e669-e678.
7. Anderson J. The MCQ controversy—a review. Med Teach. 1981; 3(4): 150-56.
8. Brassil CE, Couch BA. Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions: a Bayesian item response model comparison. Int J STEM Educ 2019; 6(1): 1-17.
9. Couch BA, Hubbard JK, Brassil CE. Multiple–true–false questions reveal the limits of the multiple–choice format for detecting students with incomplete understandings. BioScience 2018; 68(6): 455-63.
10. Simbak NB, Aung MMT, Ismail SB, et al. Comparative study of different formats of MCQs: multiple true-false and single best answer test formats, in a New Medical School of Malaysia. International Med J 2014; 21(6): 562-66.
11. McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. Med Teach 2004; 26(8): 709-12.
12. Anderson J. Medical teacher 25th anniversary series multiple-choice questions revisited. Med Teach 2004; 26(2): 110-3.
13. Gross LJ. Scoring multiple true/false tests: some considerations. Eval Health Profess 1982; 5(4): 459-68.
14. Schmidt D, Raupach T, Wiegand A, Herrmann M, Kanzow P. Relation between examinees' true knowledge and examination scores: systematic review and exemplary calculations on Multiple-True-False items. Educ Res Rev 2021; 34: 100409.
15. Sim S-M, Rasiah RI. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. Ann Acad Med Singapore2006; 35(2) :67
16. Puthiaparampil T, Rahman MM, Gudum HR, Brohi IB, Lim IF, Saimon R. How to grade items for a question bank and rank tests based on student performance. Med Ed Publish 2020; 9(1).
17. Thomas Puthiaparampil HRG, M. Mizanur Rahman RS, Lim IF. True-false analysis reveals inherent flaws in multiple true-false tests. Int J Commun Med Public Health 2019; 6(10): 4204-8.
18. Puthiaparampil T, Singmame N, Razali SBA, Lukas SB, Shee CC, Rahman MM. Dropping the non-core subjects from undergraduate final professional examination: How it would impact the results. Med J Malaysia 2022; 77(2): 169-73.
19. Burton RF. Quantifying the Effects of Chance in Multiple Choice and True/False Tests: Question selection and guessing of answers. Assess Eval Higher Educ. 2001; 26(1): 41-50.
20. Burton RF. Multiple choice and true/false tests: reliability measures and some implications of negative marking. Assess Eval Higher Educ 2004; 29(5): 585-95.